DOCUMENT RESUME

ED 427 053 TM 029 434

AUTHOR Rheinheimer, David C.; Penfield, Douglas A.
TITLE The Effects of Type I Error Rate and Power of the ANCOVA F-Test and Selected Alternatives under Non-Normality and Variance Heterogeneity.
PUB DATE 1998-10-09
NOTE 28p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, October 28-30, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Covariance; Monte Carlo Methods; *Robustness (Statistics); Sample Size; Simulation
IDENTIFIERS *F Test; Nonnormal Distributions; *Type I Errors; Variance (Statistical)

ABSTRACT

The performance of analysis of covariance (ANCOVA) and six selected competitors was examined under varying experimental conditions through Monte Carlo simulations. The six alternatives were: (1) Quade's procedure (D. Quade, 1967); (2) Puri and Sen's solution (M. Puri and P. Sen, 1969); (3) Burnett and Barr's rank difference scores (T. Burnett and D. Barr, 1977); (4) Conover and Iman's rank transformation test (W. Conover and R. Iman, 1982); (5) Hettmansperger's procedure (T. Hettmansperger, 1984); and (6) the Puri-Sen-Harwell-Serlin test (R. Harwell and R. Serlin, 198). The conditions that were manipulated included assumptions of normality and variance homogeneity, sample size, number of treatment groups, strength of the covariate/dependent variable relationship, and multiple combinations of these factors. Results indicate that variance heterogeneity, especially in combination with unbalanced designs and severe nonnormality, had a profound impact on Type I error rates. The ANCOVA F-test was robust and exhibited high power under variance homogeneity, and for some cases of variance heterogeneity, but became less competitive as conditions departed from normality. (Contains 4 tables and 23 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 427 053

The Effects on Type I Error Rate and Power of the ANCOVA F-Test and Selected Alternatives Under Non-Normality and Variance Heterogeneity

David C. Rheinheimer

East Stroudsburg University

Douglas A. Penfield

Rutgers University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Rheinheimer David

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

October 9, 1998

Paper to be presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, New York.

Correspondence concerning this paper should be directed to David Rheinheimer, 25 Rosenkrans East, East Stroudsburg University, East Stroudsburg, PA 18301.

TM029434

BEST COPY AVAILABLE

Abstract

The performance of analysis of covariance (ANCOVA) and six selected competitors was examined under varying experimental conditions through Monte Carlo simulations. The six alternatives were Quade's procedure, Puri and Sen's solution, Burnett and Barr's rank difference scores, Conover and Iman's rank transformation test, Hettmansperger's procedure, and the Puri-Sen-Harwell-Serlin test. The conditions that were manipulated included assumptions of normality and variance homogeneity, sample size, number of treatment groups, strength of the covariate/dependent variable relationship, and multiple combinations of these factors. Results indicated that variance heterogeneity, especially in combination with unbalanced designs and severe nonnormality, had a profound impact on Type I error rates. The ANCOVA F-test was robust and exhibited high power under variance homogeneity, and for some cases of variance heterogeneity, but became less competitve as conditions departed from normality.

Introduction

Next to analysis of variance, analysis of covariance may be the most popular procedure for comparing group means in educational and behavioral studies. Schneider (1996) reported that ANOVA and ANCOVA together accounted for almost 35% of the statistical techniques used in three leading educational research journals from 1978 to 1987. When the subjects under study are found to differ on one or more preexisting conditions, the analysis of covariance offers the major advantages over ANOVA of greater statistical power and a reduction in bias (Frigon and Laurencelle, 1993).

The analysis of covariance procedure combines regression analysis and analysis of variance to adjust for the effects of one or more covariates. The model for one covariate can be written as:

$$Y_{ij} = \mu + \tau_j + \beta(X_{ij} - \mu_x) + \varepsilon_{ij}$$

$$i = 1,..,N;\ j = 1,..,t,$$

where Y_{ij} is the value for the ith subject in the jth group on the dependent variable Y, μ is the grand population mean across all observations, $\tau_j = \mu_j - \mu$ is the treatment effect, β is the slope between the covariate and the dependent variable, $X_{ij} - \mu_x$ is the deviation of the covariate score about the grand X mean, and ε_{ij}

is the error term. This model can be extended to two or more covariates and to factorial designs.

Although ANCOVA is similar in its application to analysis of variance (ANOVA), the presence of covariates reduces the ANCOVA error variance and offers a more sensitive test for the hypothesis that the population means of the dependent variable do not differ. However, this greater sensitivity comes at the expense of a set of assumptions additional to those underlying traditional ANOVA. Violations of one or more of these assumptions may threaten the validity of the ANCOVA results and warrant the consideration of another test.

Review of Literature

ANCOVA Assumptions

The eight assumptions Huitema (1980) recognized as underlying proper application of fixed effects ANCOVA (randomization, homogeneity of within-group regressions, statistical independence of covariates and treatments, fixed covariate values that are error free, linearity of within-group regressions, normality of conditional Y scores, homogeneity of variance of conditional Y scores, and fixed treatment levels) include three which meet Johnson and Rakow's (1994) classification of data set violations, those concerned with the data set and parent population. These three assumptions

(linearity, normality, and homogeneity of variances) are directly a consequence of data set problems, are well-suited for Monte Carlo studies, and are necessary for statistical simplicity and validity of statistical tests (Elashoff, 1969).

Atiqullah (1964) investigated mathematically the effects of nonlinearity on the ANCOVA *F* test and reported that nonlinear regression produced a biased treatment effect. More recently, Harwell (1997) studied the effect of a nonlinear regression term on the behavior of the ANCOVA *F* test and found that the presence of a quadratic term had little effect on Type I error rates, but power was affected. His Monte Carlo simulations showed that power losses could be as high as 20% and depended on the magnitude of the nonlinear term's regression parameter.

Normality of conditional Y scores requires that the dependent variable values be normally distributed at each level of the covariate. Huitema (1980) surmised that ANCOVA may be more sensitive to departures from normality than ANOVA. He felt that Monte Carlo studies were needed on the effects of sample size, skewness, and kurtosis together to determine the degree of bias caused by conditional nonnormality of the dependent variable.

Most studies have found ANCOVA to be reasonably robust to moderate violations of the normality assumption, but when conditional nonnormality was combined with other violations the

results were less conclusive. Olejnik and Algina (1984) reported that ANCOVA tended to be conservative when conditional nonnormality was combined with heteroscedasticity, small sample sizes, and nominal α equal to 0.05. Seaman, Algina, and Olejnik (1985) found power advantages for ANCOVA over the alternatives tested except when the correlation of the sign of the skew and effect size was negative, in which case the power differences were small.

Conover and Iman (1982) compared parametric ANCOVA with rank ANCOVA procedures for four nonnormal distributions: (a) lognormal, (b) exponential, (c) uniform, and (d) Cauchy. Their results showed parametric ANCOVA to be conservative when applied to the lognormal and Cauchy distributions, and reported power advantages for the distribution-free approaches when the conditional distributions were exponential and Cauchy.

More recently, Harwell and Serlin (1988) and Johnson and Rakow (1994) have investigated the effects on ANCOVA of a number of conditions, including conditional nonnormality of Y. Harwell and Serlin matched four conditional Y distributions (normal, double exponential, exponential, and approximate Cauchy) with equal and unequal treatment group slopes, and equal and unequal group sample sizes for power and Type I error analyses. They found that the parametric *F* test maintained good Type I error rates across a variety of nonnormal distributions and other

simulation conditions for both equal and unequal slopes, but the power advantage for nonparametric tests expanded with increasing nonnormality irregardless of slope conditions. Johnson and Rakow explored the effects of unequal sample sizes, unequal group regression slopes, and group variance heterogeneity on the robustness of ANCOVA and included a range of shape perturbations. They found that combinations of unequal group variances, sample sizes, and regression slopes posed the greatest threat to ANCOVA robustness, but the authors did not extend their research to power considerations.

The assumption of homogeneity of variance of conditional *Y* scores has two cases in which a violation may occur: (a) the variance of the conditional *Y* scores is assumed to be the same for each treatment group, and (b) the variance of the conditional *Y* scores should not depend on the value of *X* (heteroscedasticity). The first case, different treatment group variances on *Y* but constant within groups variance across *X*, is of greatest concern when found in the presence of unbalanced designs, and perhaps with other assumption violations or sample conditions. Huitema (1980) concluded from his review of other studies that, similar to the patterns found for ANOVA, the effect of group variance and sample size differences depends on how the variance and sample sizes are associated. When the larger variances are associated with the larger sample sizes, the *F* test

is conservative, and when the variance/sample size matchings are inversely related, the bias is liberal.

Alternatives to ANCOVA

Among the most frequently cited nonparametric alternatives to ANCOVA are procedures proposed by Quade (1967), Puri and Sen (1969), McSweeney and Porter (1971), Burnett and Barr (1977), Shirley (1981), Conover and Iman (1982), Hettmansperger (1984), and Harwell and Serlin (1989). These tests have been the subjects of a number of simulation studies and reviews in which their performance has been compared to that of parametric ANCOVA (Olejnik & Algina, 1984; Olejnik & Algina, 1985; Seaman, Algina, & Olejnik, 1985; Harwell & Serlin, 1988).

Although Monte Carlo studies have been constructed to investigate the performance of ANCOVA and its alternatives, the published research remains limited in both the extent and depth of experimental conditions and alternatives considered. Power studies are underrepresented in the literature (Olejnik & Algina, 1984), and few studies have included both a wide range of simulation conditions and more than one or two alternatives. Most studies have restricted their range of assumption violations and sample conditions, or the number of alternatives, or both. The most extensive study found (Harwell and Serlin, 1988) included four simulation factors, but did not consider variance

heterogeneity, was limited to three nonnormal distributions, and ran only 2,000 replications per condition. The present study assessed the robustness of parametric ANCOVA under a variety of conditions and situations, and compared its performance with six of the eight alternatives cited above: Quade's procedure, Puri and Sen's solution, Burnett and Barr's rank difference scores, Conover and Iman's rank transformation test, Hettmansperger's procedure, and the Puri-Sen-Harwell-Serlin test.

Methodology

Simulation Design

Hoaglin and Andrews (1975) suggested that Monte Carlo studies be treated as statistical sampling experiments, such as factorial designs with crossed factors. With this approach, the effects under study become the factors that are manipulated in order to define the simulations. Such factors may include the number of groups, distributional parameters (e.g. kurtosis, skewness), and group sample sizes.

The most frequently mentioned technique for simulating nonnormal data employs a power transformation developed by Fleishman (1978). From the equation $W = -c + bz + cz^2 + dz^3$ a standard normal variable z can be transformed into a new variable W having the desired skewness and kurtosis. The constants c, b, and d are chosen accordingly from a table prepared by Fleishman. This method has gained wide-spread acceptance and has been used

by numerous researchers in ANCOVA studies (Olejnik & Algina, 1984; Seaman, Algina, & Olejnik, 1985; Harwell & Serlin, 1989; Harwell & Serlin, 1988; Harwell, 1997).

Six simulation factors were chosen for this study: (a) groups (two levels, 3 and 5), (b) strength of *X* / *Y* relationship (three levels, .2, .5, .8), (c) sample sizes (four levels for the three group case and three levels for the five group case), (d) conditional *Y* distribution (normal and four levels of nonnormal, (e) group variances for conditional *Y* scores (five levels of group variance ratios), and (f) treatment levels (null and nonnull). One covariate was used for all cases, and its distribution remained as standard normal throughout the study. Two levels of significance, $\alpha = 0.01$ and 0.05, were reported for all null tests, and power was computed for each nonnull case that maintained an acceptable α. The values of skewness and kurtosis ([s,k]) selected for this study, which were representive of ranges discussed in the literature, were: [0,0] (normal distribution), [0,1], [0,-1] (uniform distribution), [0.5,1], and [1.5,3].

For each replication two sets of standard normal random variables were generated with the SAS RANNOR function, a dependent variable was created with a designated correlation with the covariate, and the dependent variable was transformed to the desired degrees of skewness and kurtosis using Fleishman's power

transformation. The transformed data were subdivided into groups for further alteration and analysis.

Prior to the power transformation of the data, a new variable was created which was correlated to the initial variable that was generated. The initial variable became the covariate in the later analyses, while the newly created variable was the dependent variable. The correlation between the variables was one of the factors manipulated in the study.

The within-group sample sizes are an important variable in ANCOVA simulation studies because of their relationship to the power of the test and their interaction with group variance inequalities. Generally, the power advantage of parametric ANCOVA over rank ANCOVA increases as sample size increases, so smaller sample sizes should favor the rank ANCOVA tests (Huitema, 1980).

A further consideration in selecting sample sizes for a simulation study is the interaction between group sample size and variance inequality among the groups. According to Huitema (1980), the most sensitive situations for the ANCOVA *F* test occur when heterogeneous group variances exist with unbalanced designs. When variance and sample sizes differ, the direction of the differences appears to dictate the bias of the test. The sample sizes used for this study included equal and unequal sample size designs. For the equal sample size designs, 10 and 25 for both

the 3 and 5 group configurations were used. Three arrangements were used for the unbalanced designs, 5, 10, 15, and 10, 20, 30 for 3 groups, and 5, 10, 15, 20, 25 for 5 groups.

The variance ratios that were investigated in this study were 1:1:1, 1:1:4, and 4:1:1 for the 3 group designs, and 1:1:1:1:1, 1:1:4:4:4, and 4:4:4:1:1 for the 5 group division. These designs allowed for an examination of the effects of matching the largest group variance with the largest group sample size, and alternatively, matching the largest group variance with the smallest sample size. All combinations of the simulation factors were included, and each simulation was replicated 10,000 times.

The Simulation Procedure

Two streams of data were generated from the SAS RANNOR function, which uses the Box and Muller (1958) transformation to create standard normal random variables. These two random variables, X and X_1, created the dependent variable through the equation $Y = rX + X_1(1 - r^2)^{1/2}$, where r was the nominal correlation between the covariate, X, and the dependent variable, Y. For normal conditional distributions and variance homogeneity, samples generated with Y, which had mean zero and variance one, and X were used in the ANCOVA F test and its alternatives.

The performances of the seven tests were assessed at two levels of significance, $\alpha = 0.01$ and 0.05, by computing the rates of rejection for each of the procedures. The rate of rejection was the ratio of the significant results obtained to the number of replications performed. To account for sampling error associated with the estimated Type I errors, Bradley's liberal criterion, $.5\alpha \leq \alpha^* \leq 1.5\alpha$, was used to establish sampling error ranges around α. For $\alpha = 0.05$, the sampling error interval was (0.025, 0,075), and for $\alpha = 0.01$ the similarly calculated interval was (0.005, 0.015). Estimated error rates outside these intervals were considered conservative or liberal.

For non-normal conditional distributions, a new variable was created using Fleishman's procedure. The new variable still had mean zero and variance one, but the skewness and kurtosis could be altered as desired. To violate the assumption of group variance homogeneity, the dependent variable values were multiplied by their respective treatment group standard deviations. The Type I error rates were assessed for all seven test procedures under these assumption violations.

Power was investigated by further perturbing the dependent variable values through the addition of a treatment effect specific to each group. For the three group case, the treatment effects were -0.5, 0.0, and 0.5, while the five group case had

the same overall range of 1.0 standard deviation, but with the inclusion of the two intermediate values, -0.25 and 0.25. A power analysis was conducted on all tests that maintained the nominal Type I error rate within the appropriate sampling error intervals.

To confirm that the conditions of the experimental design were maintained, tests were performed on the data generated for the simulations. The nominal covariate/dependent variable correlations were checked by a routine incorporated into the simulation program that returned the actual correlations between the generated random variables.

Results

The mean Type I error rates for all test statistics are given in Tables 1 - 3, and the percentages of robust results by distributional shape are presented in Table 4. Each table contains the mean results of the simulations for the ANCOVA F-test and all six alternatives identified for this study. The values for the seven test statistics are given in the columns under the headings F (ANCOVA *F*), Q (Quade's Distribution-Free Test), PS (Puri and Sen's Solution), BB (Burnett and Barr's Rank Difference Scores), CI (Conover and Iman's Rank Transformation Procedure), H (Hettmansperger's Procedure), and PSHS (the Puri-Sen-Harwell-Serlin Test). For each cell the first row represents

the $\alpha = 0.05$ level of significance and the second row the $\alpha = 0.01$ level.

Bradley's (1978) liberal criterion was used for the sampling error range rather than 95% confidence intervals to allow more power analyses to be conducted. Power analyses were only conducted for the Type I error rates that were maintained within the $\alpha = 0.05$ sampling interval defined by Bradley's liberal criterion ($0.025 \leq \alpha^{*} \leq 0.075$). This allowed for the maximum number of simulations to be considered for power analyses.

Summary of Type I Error Rates

An examination of the results for each of the four designs under the combinations of correlation, skewness, kurtosis, and variance ratios showed that the patterns of Type I error rates were dependent primarily on the presence or absence of variance homogeneity and the degree of skewness and kurtosis, with some lesser effects from the sample size configuration. Under variance homogeneity, all tests maintained excellent control of Type I error for all sample size designs even under the harshest conditions of skew and kurtosis.

The presence of variance heterogeneity affected robustness similarly for the equal sample size designs, with little difference between the 3 and 5 group designs. The F test maintained the best control over Type I error across sample size,

number of groups, and distributional shape when variance heterogeneity was combined with equal sample sizes.

Variance heterogeneity with the unequal sample sized design was separated into two divisions, matching the largest group variance with the largest group sample size, and the inverse coupling. The results for these two categories were quite different. Under the first variance ratios (1:1:4 or 1:1:4:4:4), most tests behaved like the equal sample size designs. Across distributional shapes the unequal sample size designs performed nearly as well as the equal sample size classification, except for the most severe case of skew and kurtosis, when the unequal sample size design was much more robust for both the 3 and 5 group designs (Table 4). Otherwise, the 5 group unequal sample size designs did only slightly better than the respective 3 group designs.

When distributional shapes were combined, the unequal sample sized rank based tests outperformed their equal sample sized counterparts for both the 3 and 5 group designs. The F test performance, however, was superior when sample sizes were equal. Generally, control over Type I error did not improve when sample sizes were increased.

Matching the largest sample group variance with the smallest group size had a dramatic effect on Type I error control, causing very liberal results that were rarely ever robust. Only three

tests, BB, H, and PSHS, produced any robust results, the best showing coming from the BB test at 60% Type I error control for the n = 5, 10, 15 sample size design. The poor control over Type I error was equally evident when data were examined by distributional shape, although the 3 group, unequal sample size design did perform slightly better than its 5 group complement. Overall rates of robustness for each statistic under variance heterogeneity (for $\alpha = .05$) were 54.7% for the F test, 59.3% for the Q and PS tests, 74% for the BB test, 58.7% for the CI test, 65.3% for the H test, and 61.3% for the PSHS test. The performance was similar for all tests under variance homogeneity, with respective percentages given as 73.3, 76.1, 76.1, 84.7, 75.7, 79.6, and 77.2.

Power Analyses: Summary

The factor that had the greatest effect on power was variance heterogeneity. When sample sizes were unequal, variance heterogeneity depressed power when the largest group variance was coupled with the largest sample size, and prevented any power analyses when the variance/sample size association was indirect. The strength of the covariate/dependent variable association also affected power, causing it to increase with increasing r.

Under variance homogeneity and normal to moderately nonnormal conditions, the F test was generally the most powerful test. As

conditions became more nonnormal, the advantage shifted to the Q and CI tests, which were (usually) most powerful. The BB test was consistently the least powerful statistic.

The F test continued to exhibit good power under variance heterogeneity when the sample sizes were equal and conditions favored normality. As nonnormality increased, the alternatives to the F test became more powerful, with the Q or CI tests generally most powerful. With unequal sample sizes and the largest treatment group variance and sample size directly matched, the CI and Q tests were again the most powerful, while the F test had generally lowest power.

Overall, no test was universally most powerful, but the F, Q, and CI tests were more frequently the most powerful. Under variance homogeneity and for most distributional shapes, or when sample sizes were equal and variance heterogeneity was present, the F test was as powerful, or more powerful, than any other test. When nonnormality was most severe, or when sample sizes were unequal and variance heterogeneity with direct variance ratio/sample size coupling was present, the Q and CI tests were most powerful, followed closely by the PS test. Under variance heterogeneity, unequal sample sizes, and the inverse matching of variance ratio and sample size, no test was robust enough to be most powerful.

Conclusions

Both Type I error control and power were affected greatly by variance heterogeneity. Under variance homogeneity, all tests maintained excellent control of Type I error for all sample size designs even under the harshest conditions of skew and kurtosis. When variance heterogeneity was coupled with unbalanced designs, such that the largest treatment group variance was matched with the largest group sample size, the nonparametric alternatives, especially the Conover and Iman and Quade's procedures, were most robust and had highest power. When variance heterogeneity was combined with the inverse coupling of sample size and variance ratio, no test maintained adequate control over Type I error. The strength of the covariate/ dependent relationship had a pronounced effect on power, causing it to decrease as the relationship weakened.

References

Atiquallah, M. (1964). The robustness of the covariance analysis of a one-way classification. Biometrika, 51, 365-372.

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. Annals of Mathematical Statistics, 28, 610-611.

Bradley, J. V., (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Burnett, T. D., & Barr, D. R. (1977). A nonparametric analogy of analysis of covariance. Educational and Psychological Measurement, 37, 341-348.

Conover, W. J., & Iman, R. L. (1982). Analysis of covariance using the rank transformation. Biometrics, 38, 715-724.

Elashoff, J. A. (1969). Analysis of covariance: A delicate instrument. American Educational Research Journal, 6,(3), 383-401.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. Psychometrika, 43,(4), 521-532.

Frigon, J. Y., & Laurencelle, L. (1993). Analysis of covariance: A proposed algorithm. Educational and Psychological Measurement, 53(1), 1-18.

Harwell, M. R., & Serlin, R. C. (1988). An empirical study of a proposed test of nonparametric analysis of covariance.

Psychological Bulletin, 104(2), 268-281.

Harwell, M. R., & Serlin, R. C. (1989). A nonparametric test statistic for the general linear model. Journal of Educational Statistics, 14(4), 351-371.

Harwell, M. (1997). An investigation of the effect of nonlinearity of regression on the ANCOVA F test. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Hettmansperger, T. P. (1984). Statistical inference based on ranks. New York: John Wiley.

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. The American Statistician, 29, 122-126.

Huitema, B. E. (1980). The analysis of covariance and alternatives. New York: John Wiley.

Johnson, C. C., & Rakow, E. A. (1994). Effects of violations of data set assumptions when using the analysis of variance and covariance with unequal group sizes. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Nashville, TN.

McSweeney, M., & Porter, A. C. (1971). Small sample properties of nonparametric index of response and rank analysis of covariance. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Olejnik, S. F., & Algina, J. (1985). A review of nonparametric alternatives to analysis of covariance. Evaluation Review, 9(1), 51-83.

Olejnik, S. F., & Algina, J. (1984). Parametric ANCOVA and the rank transform ANCOVA when the data are conditionally non-normal and heteroscedastic. Journal of Educational Statistics, 9(2), 129-149.

Puri, M. L., & Sen, P. K. (1969). Analysis of covariance based on general rank scores. Annals of Mathematical Statistics, 40(2), 610-618.

Quade, D. (1967). Rank analysis of covariance. Journal of the American Statistical Association, 62, 1187-1200.

Schneider, P. J. (1996). The effects on Type I error rate and power of the ANOVA F-test and selected alternatives under non-normality and variance heterogeneity (Doctoral dissertation, Rutgers University, 1996). Dissertation Abstracts International, 57(05), 3280B.

Seaman, S. L., Algina, J., & Olejnik, S. F. (1985). Type I error probabilities and power of the rank and parametric ANCOVA procedures. Journal of Educational Statistics, 10(4), 345-367.

Shirley, E. A. C. (1981). A distribution-free method for analysis of covariance based in ranked data. Applied Statistics, 30(2), 158-162.

Table 1

Mean Type I Error Rates at $\alpha = 0.05, 0.01$, under variance homogeneity.

3 Groups

n	α	F	Q	PS	BB	CI	H	PSHS
10	0.05	.0495	.0506	.0501	.0488	.0507	.0418	.0446
	0.01	.0098	.0108	.0078	.0097	.0109	.0056	.0062
25	0.05	.0498	.0500	.0499	.0491	.0500	.0464	.0478
	0.01	.0092	.0102	.0090	.0099	.0103	.0078	.0083
5[a]	0.05	.0499	.0509	.0503	.0497	.0513	.0403	.0450
	0.01	.0102	.0103	.0071	.0095	.0102	.0053	.0059
10[b]	0.05	.0490	.0498	.0497	.0500	.0498	.0455	.0472
	0.01	.0101	.0102	.0088	.0092	.0102	.0075	.0081

5 Groups

n	α	F	Q	PS	BB	CI	H	PSHS
10	0.05	.0501	.0509	.0475	.0492	.0510	.0409	.0434
	0.01	.0100	.0106	.0071	.0098	.0105	.0055	.0061
25	0.05	.0506	.0508	.0496	.0497	.0511	.0462	.0478
	0.01	.0103	.0104	.0091	.0097	.0104	.0082	.0086
5[c]	0.05	.0491	.0501	.0476	.0497	.0498	.0419	.0448
	0.01	.0101	.0100	.0078	.0099	.0100	.0063	.0070

[a] Sample sizes were 5, 10, 15.

[b] Sample sizes were 10, 20, 30.

[c] Sample sizes were 5, 10, 15, 20, 25.

Table 2

Mean Type I Error Rates at $\alpha = 0.05, 0.01$, under variance heterogeneity, $\sigma^2 = 1:1:4$ (3 groups) and $\sigma^2 = 1:1:4:4:4$ (5 groups).

3 Groups

n	α	F	Q	PS	BB	CI	H	PSHS
10	0.05	.0661	.0762	.0755	.0674	.0764	.0555	.0688
	0.01	.0195	.0227	.0177	.0163	.0229	.0112	.0152
25	0.05	.0634	.0955	.0954	.0841	.0955	.0727	.0927
	0.01	.0181	.0237	.0311	.0250	.0334	.0199	.0286
5[a]	0.05	.0255	.0425	.0421	.0483	.0426	.0296	.0375
	0.01	.0048	.0088	.0063	.0092	.0090	.0040	.0049
10[b]	0.05	.0228	.0462	.0461	.0509	.0462	.0368	.0439
	0.01	.0041	.0105	.0092	.0109	.0105	.0063	.0086

5 Groups

n	α	F	Q	PS	BB	CI	H	PSHS
10	0.05	.0608	.0781	.0733	.0668	.0735	.0532	.0681
	0.01	.0152	.0212	.0156	.0158	.0215	.0096	.0138
25	0.05	.0596	.1053	.1034	.0908	.1053	.0696	.1009
	0.01	.0156	.0354	.0323	.0277	.0354	.0170	.0313
5[c]	0.05	.0283	.0611	.0419	.0531	.0441	.0308	.0398
	0.01	.0054	.0093	.0075	.0117	.0094	.0048	.0071

[a] Sample sizes were 5, 10, 15.

[b] Sample sizes were 10, 20, 30.

[c] Sample sizes were 5, 10, 15, 20, 25.

Table 3

Mean Type I Error Rates at $\alpha = 0.05, 0.01$, under variance heterogeneity, $\sigma^2 = 4:1:1$ (3 groups) and $\sigma^2 = 4:4:4:1:1$ (5 groups).

3 Groups

n	α	F	Q	PS	BB	CI	H	PSHS
5[a]	0.05	.1484	.0957	.0948	.0734	.0961	.0774	.0874
	0.01	.0608	.0308	.0243	.0187	.0312	.0182	.0208
10[b]	0.05	.1478	.0988	.0987	.0784	.0990	.0867	.0950
	0.01	.0628	.0325	.0297	.0209	.0326	.0242	.0281

5 Groups

n	α	F	Q	PS	BB	CI	H	PSHS
5[c]	0.05	.1397	.1343	.1301	.0984	.1343	.1006	.1252
	0.01	.0512	.0468	.0402	.0291	.0469	.0259	.0381

[a] Sample sizes were 5, 10, 15.

[b] Sample sizes were 10, 20, 30.

[c] Sample sizes were 5, 10, 15, 20, 25.

Table 4

Percentages of Robust Results by Distributional Shape Based on Bradley's Liberal Criterion at $\alpha = 0.05$.

σ^2 = 1:1:1 or 1:1:1:1:1

s,k	3 x e	5 x e	3 x u	5 x u
0,0	100.0	100.0	100.0	100.0
0,1	100.0	100.0	100.0	100.0
0,-1	100.0	100.0	100.0	100.0
.5,1	100.0	100.0	100.0	100.0
1.5,3	100.0	100.0	100.0	100.0

σ^2 = 1:1:4 or 1:1:4:4:4

s,k	3 x e	5 x e	3 x u	5 x u
0,0	100.0	100.0	90.5	95.2
0,1	100.0	100.0	90.5	95.2
0,-1	90.5	100.0	88.1	95.2
.5,1	100.0	88.1	92.9	95.2
1.5,3	11.9	16.7	83.3	71.4

(table continues)

σ^2 = 4:1:1 or 4:4:4:1:1

s,k	3 x u	5 x u
0,0	14.3	4.8
0,1	16.7	9.5
0,-1	9.5	4.8
.5,1	14.3	4.8
1.5,3	2.4	0.0

e = equal sample size grouping, u = unequal sample size grouping.